ENV 790.30 - Time Series Analysis for Energy Data | Spring 2021

Assignment 2 - Due date 01/26/22

Tatiana Sokolova

Submission Instructions

You should open the .rmd file corresponding to this assignment on RStudio. The file is available on our class repository on Github.

Once you have the file open on your local machine the first thing you will do is change "Student Name" on line 4 with your name. Then you will start working through the assignment by **creating code and output** that answer each question. Be sure to use this assignment document. Your report should contain the answer to each question and any plots/tables you obtained (when applicable).

When you have completed the assignment, **Knit** the text and code into a single PDF file. Rename the pdf file such that it includes your first and last name (e.g., "LuanaLima_TSA_A02_Sp22.Rmd"). Submit this pdf using Sakai.

R packages

R packages needed for this assignment:"forecast","tseries", and "dplyr". Install these packages, if you haven't done yet. Do not forget to load them before running your script, since they are NOT default packages.\

Data set information

Consider the data provided in the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source on our **Data** folder. The data comes from the US Energy Information and Administration and corresponds to the January 2022 Monthly Energy Review. The spreadsheet is ready to be used. Use the command *read.table()* to import the data in R or *panda.read_excel()* in Python (note that you will need to import pandas package). }

```
# Importing data set
energy_data <- read.xlsx(file = "../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Sour
    header = FALSE, startRow = 13, sheetIndex = 1)
# extracting column names from row 11
read_col_names <- read.xlsx(file = "../Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_S
    header = FALSE, startRow = 11, endRow = 11, sheetIndex = 1)
colnames(energy_data) <- read_col_names
head(energy_data)
```

```
##        Month Wood Energy Production Biofuels Production
## 1 1973-01-01                129.630       Not Available
## 2 1973-02-01                117.194       Not Available
## 3 1973-03-01                129.763       Not Available
## 4 1973-04-01                125.462       Not Available
```

```
## 5 1973-05-01                129.624       Not Available
## 6 1973-06-01                125.435       Not Available
##   Total Biomass Energy Production Total Renewable Energy Production
## 1                         129.787                            403.981
## 2                         117.338                            360.900
## 3                         129.938                            400.161
## 4                         125.636                            380.470
## 5                         129.834                            392.141
## 6                         125.611                            377.232
##   Hydroelectric Power Consumption Geothermal Energy Consumption
## 1                         272.703                         1.491
## 2                         242.199                         1.363
## 3                         268.810                         1.412
## 4                         253.185                         1.649
## 5                         260.770                         1.537
## 6                         249.859                         1.763
##   Solar Energy Consumption Wind Energy Consumption Wood Energy Consumption
## 1            Not Available          Not Available                 129.630
## 2            Not Available          Not Available                 117.194
## 3            Not Available          Not Available                 129.763
## 4            Not Available          Not Available                 125.462
## 5            Not Available          Not Available                 129.624
## 6            Not Available          Not Available                 125.435
##   Waste Energy Consumption Biofuels Consumption
## 1                    0.157        Not Available
## 2                    0.144        Not Available
## 3                    0.176        Not Available
## 4                    0.174        Not Available
## 5                    0.210        Not Available
## 6                    0.176        Not Available
##   Total Biomass Energy Consumption Total Renewable Energy Consumption
## 1                         129.787                            403.981
## 2                         117.338                            360.900
## 3                         129.938                            400.161
## 4                         125.636                            380.470
## 5                         129.834                            392.141
## 6                         125.611                            377.232
```

Question 1

You will work only with the following columns: Total Biomass Energy Production, Total Renewable Energy Production, Hydroelectric Power Consumption. Create a data frame structure with these three time series only. Use the command head() to verify your data.

```
df <- energy_data[, c("Total Biomass Energy Production", "Total Renewable Energy Production",
    "Hydroelectric Power Consumption")]
head(df)
```

```
##   Total Biomass Energy Production Total Renewable Energy Production
## 1                         129.787                            403.981
## 2                         117.338                            360.900
## 3                         129.938                            400.161
## 4                         125.636                            380.470
```

```
## 5                          129.834                                392.141
## 6                          125.611                                377.232
##   Hydroelectric Power Consumption
## 1                          272.703
## 2                          242.199
## 3                          268.810
## 4                          253.185
## 5                          260.770
## 6                          249.859
```

Question 2

Transform your data frame in a time series object and specify the starting point and frequency of the time series using the function ts().

```
ts_energy_data <- ts(data = df, start = 1, end = 585, frequency = 1)
# ts_energy_data not running the above because it takes up a bunch of pages
```

Question 3

Compute mean and standard deviation for these three series.

```
mean(df$"Total Biomass Energy Production")
```

```
## [1] 273.7839
```

```
mean(df$"Total Renewable Energy Production")
```

```
## [1] 581.1708
```

```
mean(df$"Hydroelectric Power Consumption")
```

```
## [1] 235.9653
```

```
sd(df$"Total Biomass Energy Production")
```

```
## [1] 89.42852
```

```
sd(df$"Total Renewable Energy Production")
```

```
## [1] 177.5607
```

```
sd(df$"Hydroelectric Power Consumption")
```

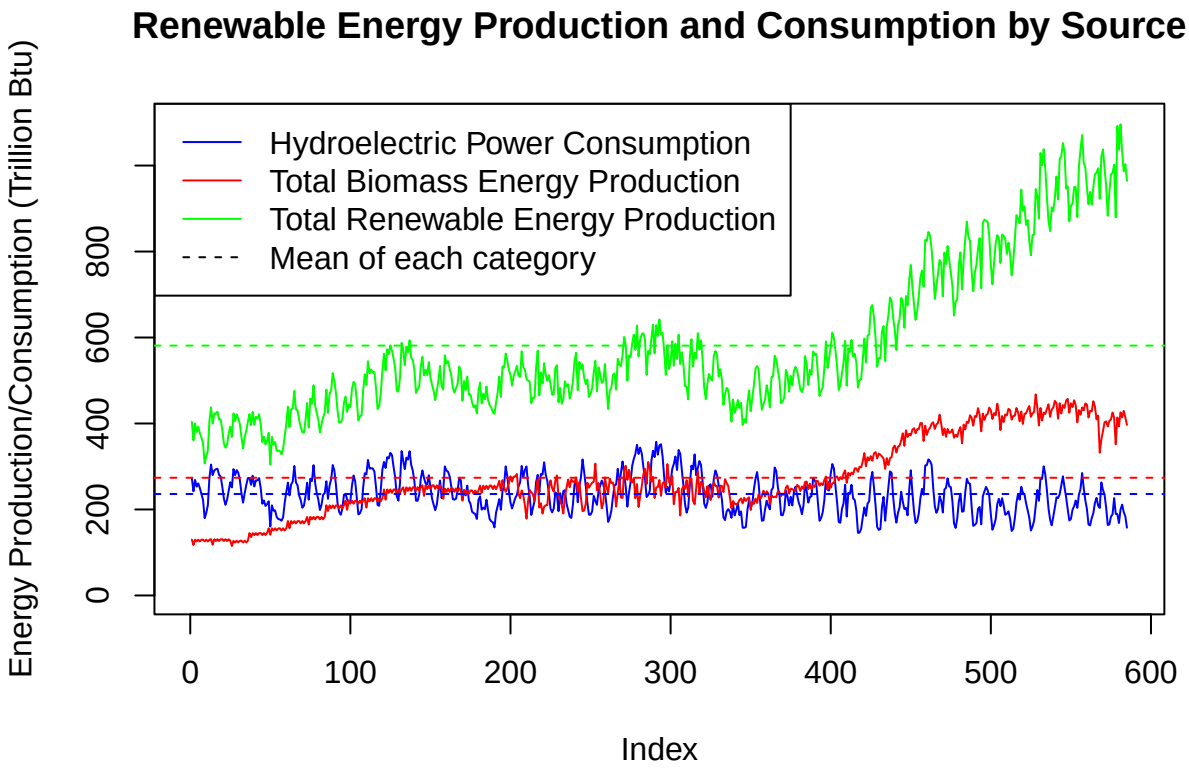
```
## [1] 44.01749
```

Question 4

Display and interpret the time series plot for each of these variables. Try to make your plot as informative as possible by writing titles, labels, etc. For each plot add a horizontal line at the mean of each series in a different color.

```
plot(df[, "Hydroelectric Power Consumption"], type = "l", col = "blue", ylab = "Energy Production/Consu
    ylim = c(0, 1100))
lines(df[, "Total Biomass Energy Production"], col = "red")
lines(df[, "Total Renewable Energy Production"], col = "green")
title(main = "Renewable Energy Production and Consumption by Source")
abline(h = mean(df[, "Hydroelectric Power Consumption"]), col = "blue", lty = "dashed")
abline(h = mean(df[, "Total Biomass Energy Production"]), col = "red", lty = "dashed")
abline(h = mean(df[, "Total Renewable Energy Production"]), col = "green", lty = "dashed")

# legend
legend("topleft", legend = c("Hydroelectric Power Consumption", "Total Biomass Energy Production",
    "Total Renewable Energy Production", "Mean of each category"), lty = c("solid",
    "solid", "solid", "dashed"), col = c("blue", "red", "green", "black"))
```



Per the time series plot, there has been a significant increase in total renewable energy production and a less rapid increase in total biomass energy production. Total biomass production has remained relatively constant. The average of total biomass energy production is slightly larger than hydroelectric power consumption's average. Whereas, as expected, the average for Total Renewable Energy production is the highest.

Question 5

Compute the correlation between these three series. Are they significantly correlated? Explain your answer.

```
cor(ts_energy_data, use = "all.obs", method = "pearson")  #linear
```

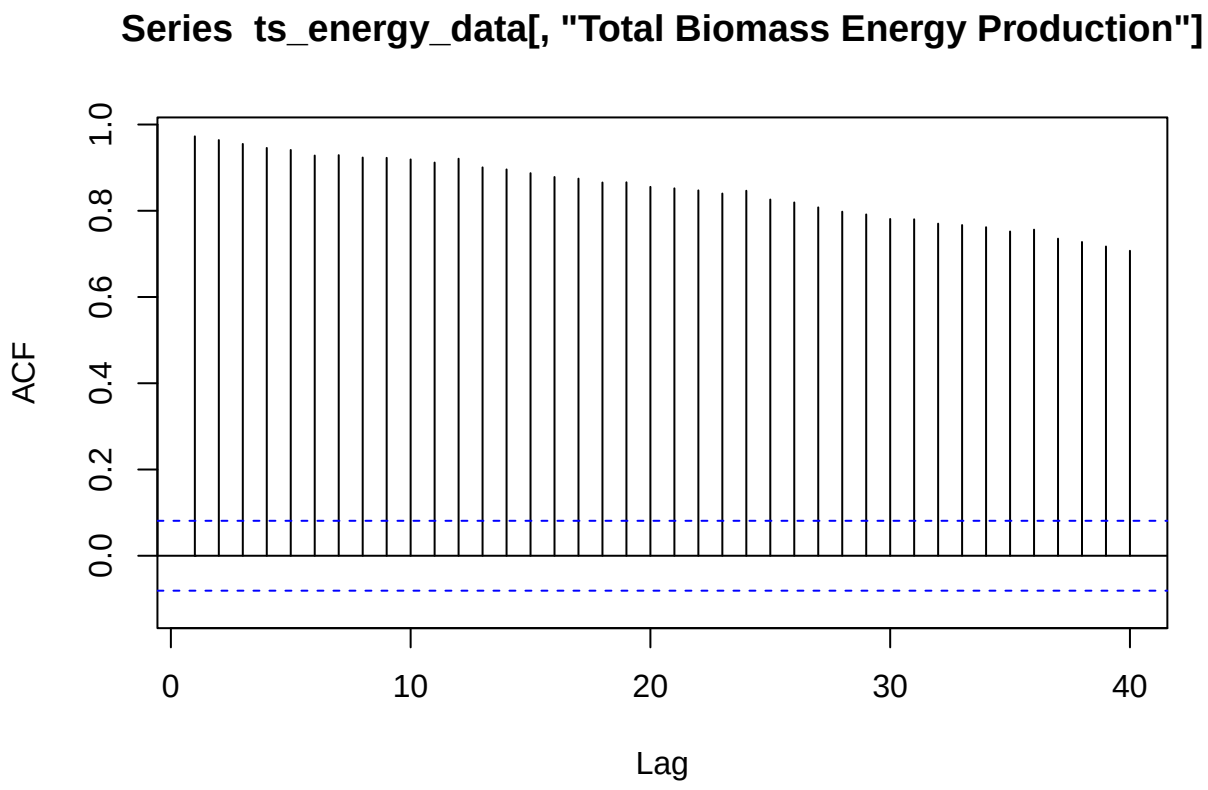
```
##                                    Total Biomass Energy Production
## Total Biomass Energy Production                          1.0000000
## Total Renewable Energy Production                        0.9232838
## Hydroelectric Power Consumption                         -0.2804997
##                                    Total Renewable Energy Production
## Total Biomass Energy Production                           0.92328377
## Total Renewable Energy Production                         1.00000000
## Hydroelectric Power Consumption                          -0.05680651
##                                    Hydroelectric Power Consumption
## Total Biomass Energy Production                         -0.28049970
## Total Renewable Energy Production                       -0.05680651
## Hydroelectric Power Consumption                          1.00000000
```

Total Renewable Energy Production and Total Biomass Energy Production have a very high positive correlation since it is above 0.9. Hydroelectric Power Consumption has negligible correlation to the other two as both correlations are below a negative .3 correlation.

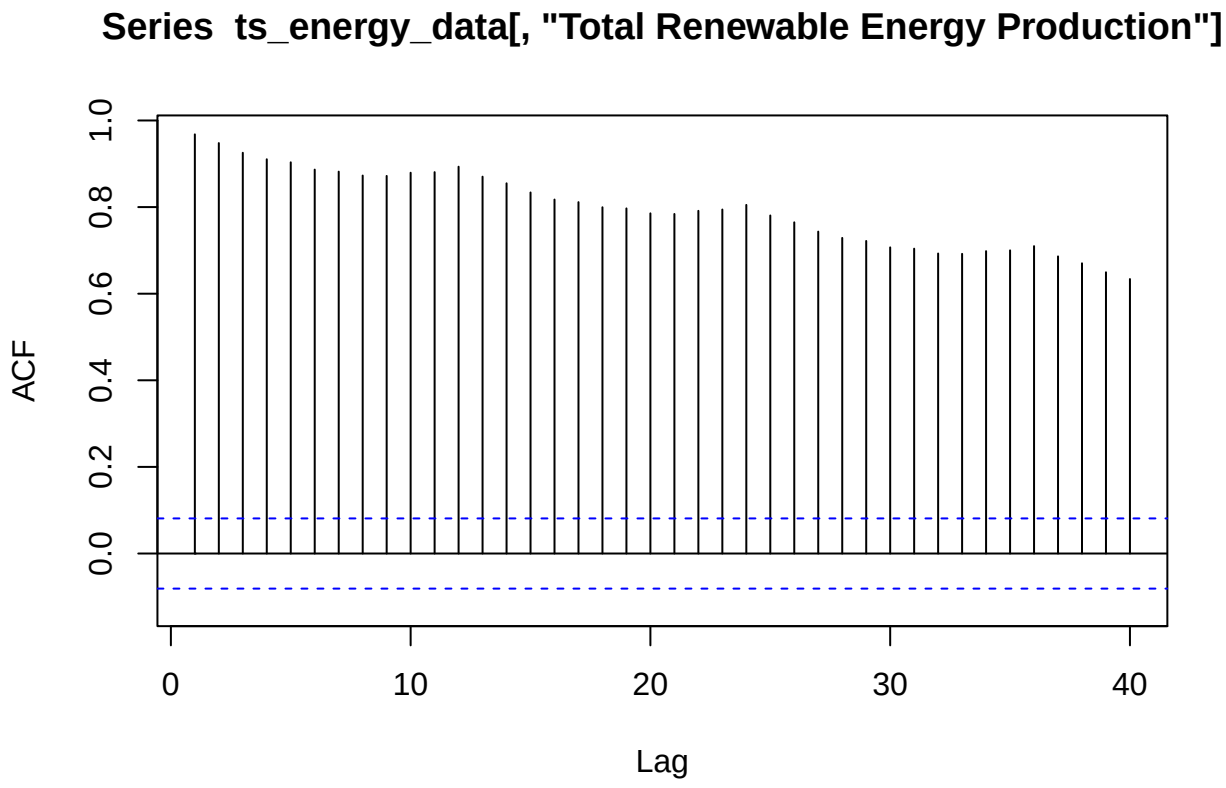
Question 6

Compute the autocorrelation function from lag 1 up to lag 40 for these three variables. What can you say about these plots? Do the three of them have the same behavior?

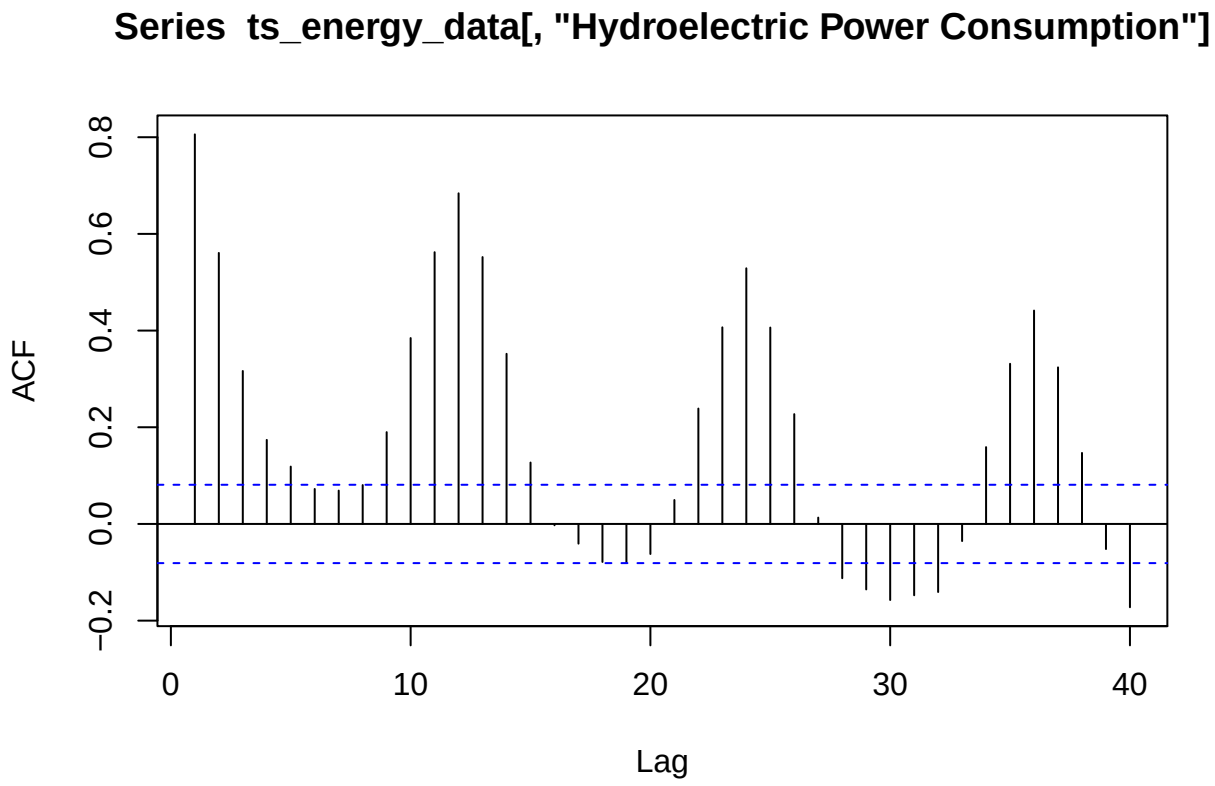
```
Biomass_acf = Acf(ts_energy_data[, "Total Biomass Energy Production"], lag.max = 40,
    type = "correlation", plot = TRUE)
```




```
RE_acf = Acf(ts_energy_data[, "Total Renewable Energy Production"], lag.max = 40,
    type = "correlation", plot = TRUE)
```




```
Hydro_acf = Acf(ts_energy_data[, "Hydroelectric Power Consumption"], lag.max = 40,
    type = "correlation", plot = TRUE)
```

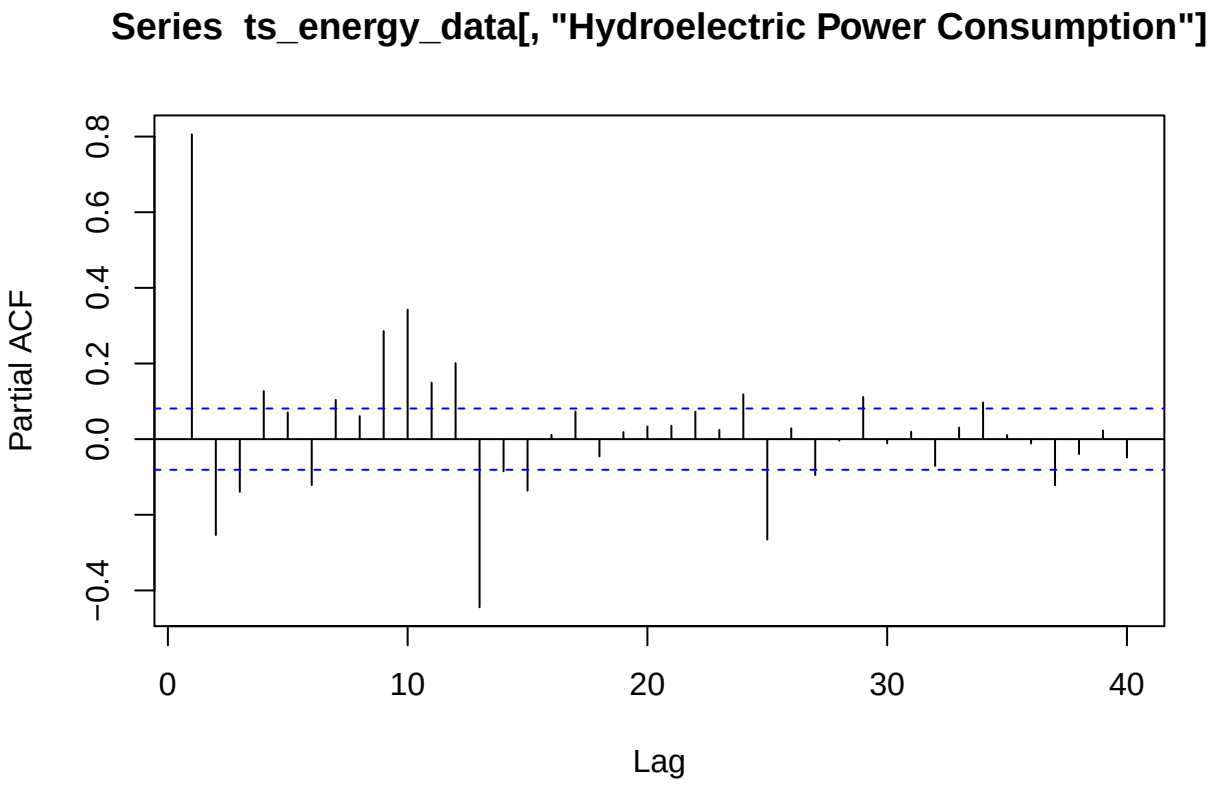


These three variables do not have the same behavior. Hydroelectric Power Consumption has seasonality. Whereas, Total Biomass Energy Production and Total Renewable Energy Production are both non-stationary since the ACF declines over time with the Total Renewable Energy Plot declining more rapidly than the Total Biomass Energy production ACF.

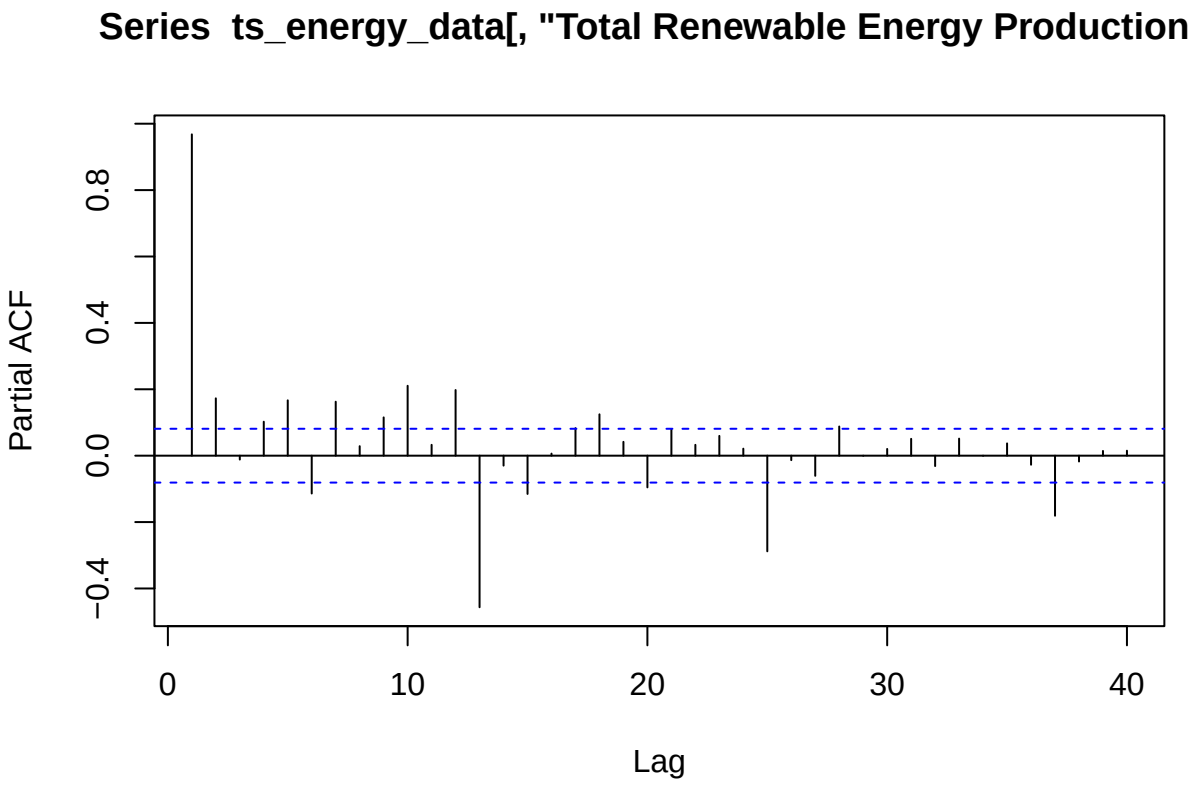
Question 7

Compute the partial autocorrelation function from lag 1 to lag 40 for these three variables. How do these plots differ from the ones in Q6?

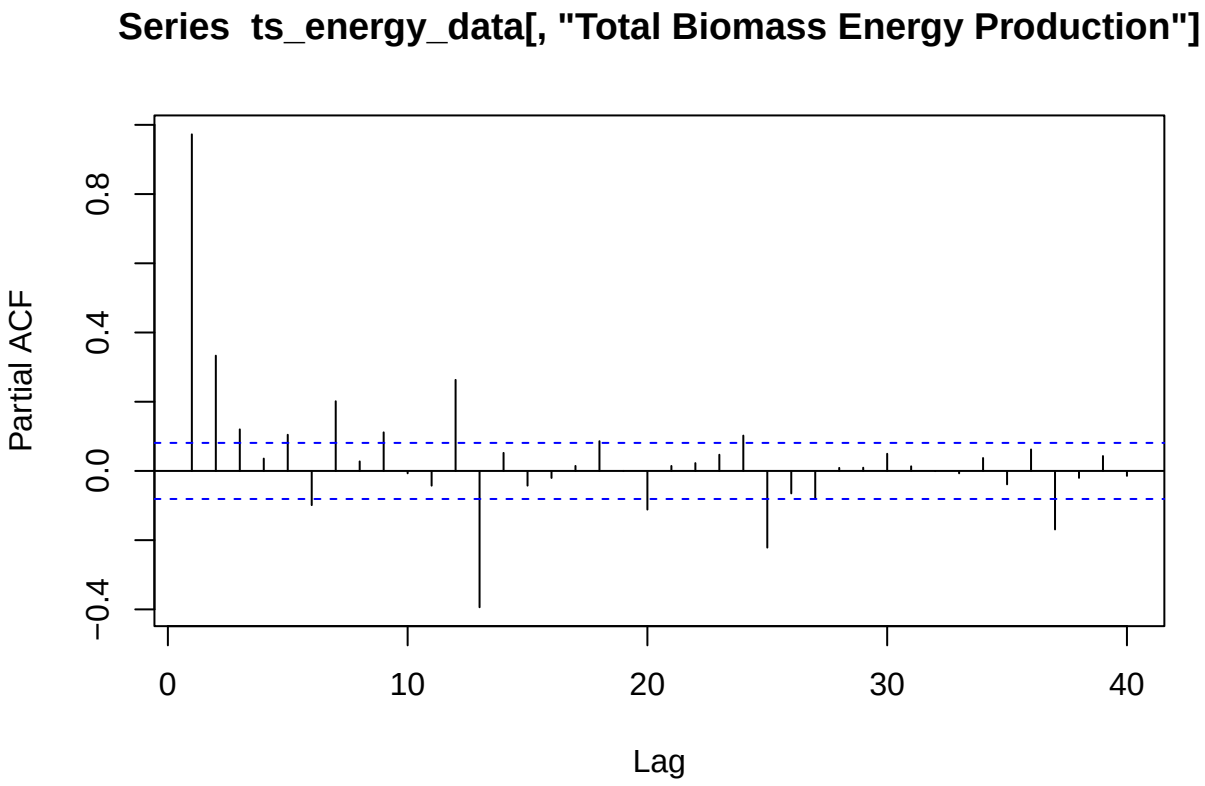
```
Hydro_pacf = Pacf(ts_energy_data[, "Hydroelectric Power Consumption"], lag.max = 40,
    plot = TRUE)
```




```
RE_pacf = Pacf(ts_energy_data[, "Total Renewable Energy Production"], lag.max = 40,
    plot = TRUE)
```




```
Biomass_pacf = Pacf(ts_energy_data[, "Total Biomass Energy Production"], lag.max = 40,
    plot = TRUE)
```



The PACF shows the partial correlation of a time series' own lagged values. Total Biomass Energy Production appears to have the most seasonality although there appears to be some seasonality for the other two variables as well.